Zgłoszenie tematu pracy dyplomowej **:: STUDIA II STOPNIA ::**

na rok akademicki 2021/22

Promotor:	**dr hab. Jozef Kapusta, prof. UP**
Temat pracy magisterskiej (j. polski oraz j. angielski):	Searching Text Similarities in Web Content *Podobieństwa tekstowe w treści internetowej*
Zakres i oczekiwane rezultaty pracy:	Document similarity (or distance between documents) is one of the central themes in Information Retrieval. How humans usually define how similar document look like? Usually, documents are treated as similar if they are semantically close and describe similar concepts. On the other hand, “similarity” can be used in the context of duplicate detection. One of the main problems of "large" web portals with many publishers is duplicity in the content. Counting the similarities of the documents could effectively help administrators of portals to optimize the web content. In the theoretical part: The theoretical part of the thesis will summarize the methods for word embedding (Word2vec, Tf-Idf, GloVe, etc.), metrics for comparison vectors of documents, techniques for web crawler etc. An important part of the work is to explore models and approaches other researchers have already developed. In the practical part: The aim of the thesis is creating a simple tool for web portals administrators. The tool will calculate and show pages from the web portal with similar content. The student will choose an appropriate document model (TF-IDF, TF, vector, boolean document model, etc.) for indexing web content, and create crawler for web portal indexing. The main task will be to calculate documents similarities and show the result for the user. An administrator could change the structure of web pages or rewrite web content, based on results from this tool.
*Aspekt naukowy, problemowy pracy:	definition and implementation of a documents model, creation method for web crawler, implementation of a method for calculating document similarities.
**Oprogramowanie, język programowania, środowisko systemowe:	Jupyter Notebook Environment (Python)
**Środowisko uruchomieniowe:	Windows or Linux
Dodatkowe wymagania i uwagi:	English language
**Literatura:	• Bird, S., Klein E., and Loper, E. (2009). Natural Language Processing with Python - Analyzing Text with the Natural Language Toolkit. O'Reilly Media.

INSTYTUT INFORMATYKI

	• Bengfort, B., Ojeda, T., Bilbro, R. (2018). Applied Text Analysis with Python: Enabling Language - Aware Data Products with Machine Learning, O'Reilly Media, 332 p. • Natural Language Toolkit, online: https://www.nltk.org/

* Regulamin studiów § 35 2. Praca dyplomowa na profilu praktycznym, podobnie jak praca inżynierska, powinna mieć charakter aplikacyjny, badawczy, projektowy lub oceniający praktykę w świetle teorii.
** pola opcjonalne